Capital Metro

Jason Antal

2024-08-18

The file capmetro_UT.csv contains data from Austin's own Capital Metro bus network, including shuttles to, from, and around the UT campus. These data track ridership on buses in the UT area. Ridership is measured by an optical scanner that counts how many people embark and alight the bus at each stop. Each row in the data set corresponds to a 15-minute period between the hours of 6 AM and 10 PM, each and every day, from September through November 2018. The variables are:

```
timestamp: the beginning of the 15-minute window for that row of data
boarding: how many people got on board any Capital Metro bus on the UT campus in the specific 15 minute
alighting: how many people got off ("alit") any Capital Metro bus on the UT campus in the specific 15 mi
day_of_week and weekend: Monday, Tuesday, etc, as well as an indicator for whether it's a weekend.
temperature: temperature at that time in degrees F
hour_of_day: on 24-hour time, so 6 for 6 AM, 13 for 1 PM, 14 for 2 PM, etc.
month: July through December
```

Your task is to create a figure, or set of related figures, that tell an interesting story about Capital Metro ridership patterns around the UT-Austin campus during the semester in question. Provide a clear annotation/caption for each figure, but the figure(s) should be more or less stand-alone, in that you shouldn't need many, many paragraphs to convey its meaning. Rather, the figure together with a concise caption should speak for itself as far as possible.

You have broad freedom to look at any variables you'd like here – try to find that sweet spot where you're showing genuinely interesting relationships among more than just two variables, but where the resulting figure or set of figures doesn't become overwhelming/confusing. (Faceting/panel plots might be especially useful here.)

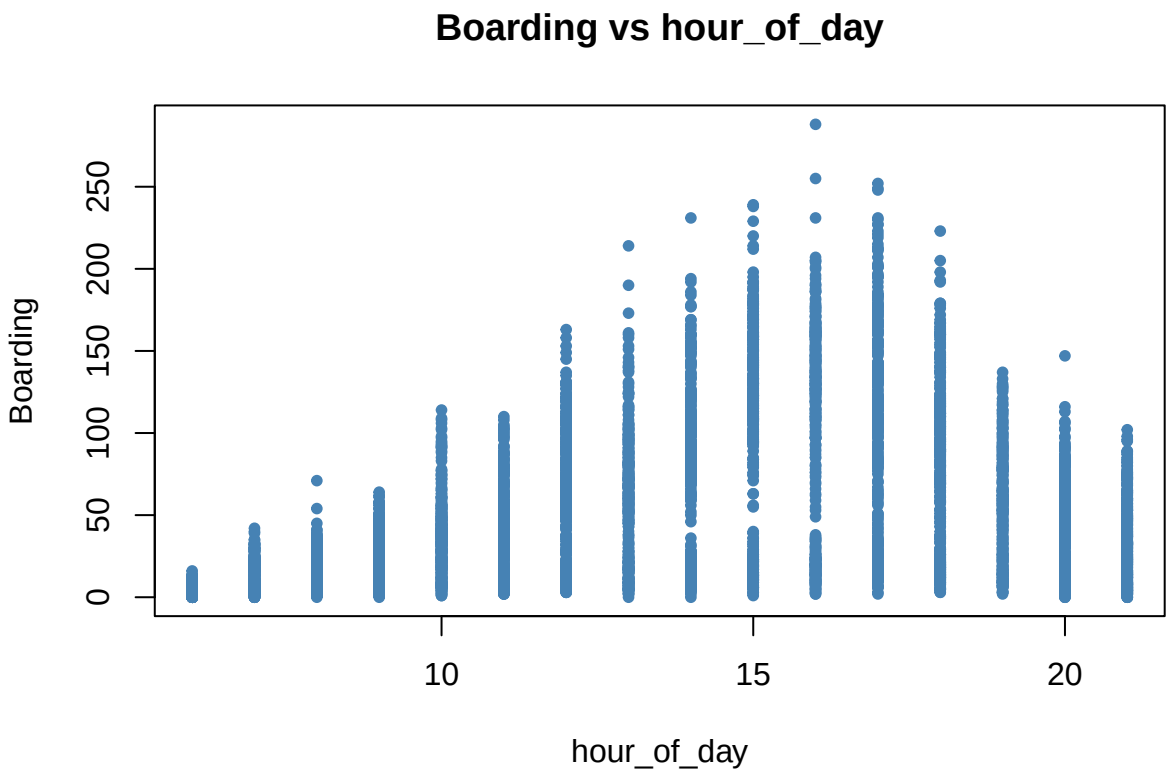
```
cat("Let's analyze all chronological data related to UT austin metro ridership
    to see what the busiest hours are. For this, we will use simple pairwise
    plotting and then perform ANOVA to determine if the relationship is signficant.")
```

```
## Let's analyze all chronological data related to UT austin metro ridership
##     to see what the busiest hours are. For this, we will use simple pairwise
##     plotting and then perform ANOVA to determine if the relationship is signficant.
```

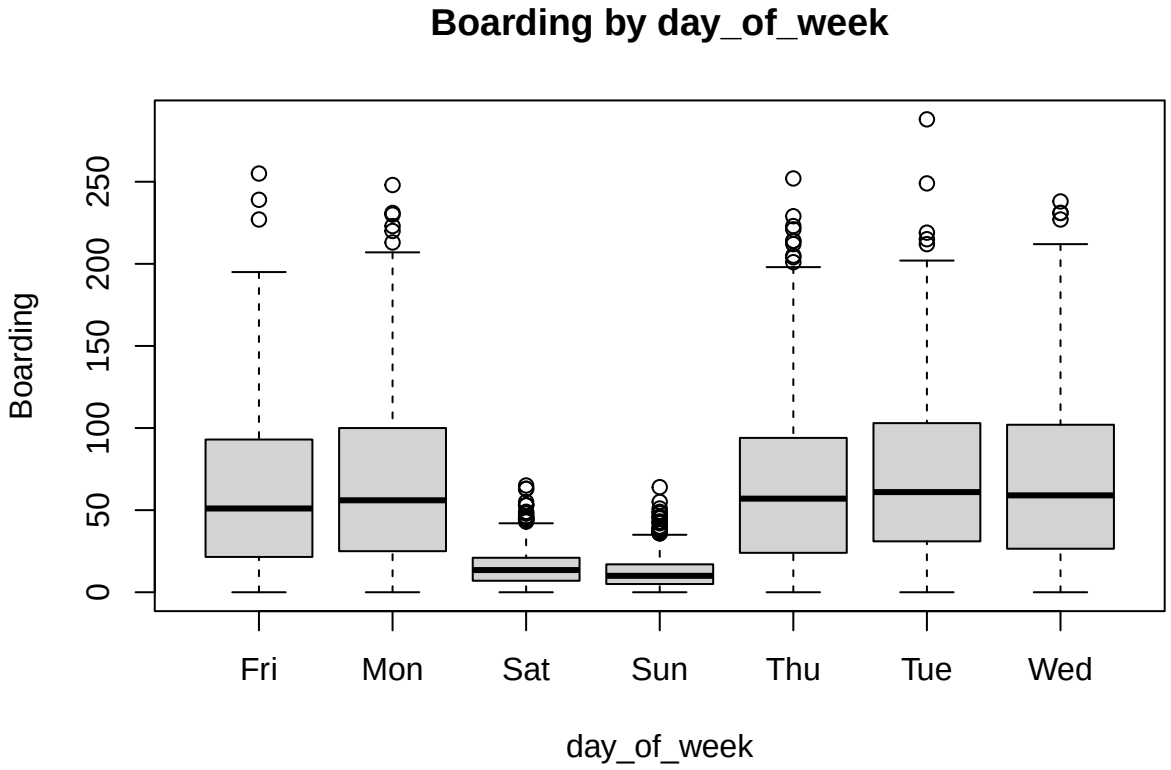
```
#pairwise analysis of all day-related data to see each chronological relationship
pairwise_analysis <- function(data) {
  variables <- c("hour_of_day", "day_of_week", "weekend")

  # Loop through each variable
  for (var in variables) {
    cat("\nAnalyzing:", var, "\n")
```

```r
    if (is.numeric(data[[var]])) {
      correlation <- cor(data$boarding, data[[var]], use = "complete.obs")
      cat("Correlation with boarding:", round(correlation, 3), "\n")

      plot(data[[var]], data$boarding,
           main = paste("Boarding vs", var),
           xlab = var, ylab = "Boarding",
           pch = 20, col = "steelblue")
    } else {
      # For categorical variables, create a boxplot
      boxplot(boarding ~ data[[var]], data = data,
              main = paste("Boarding by", var),
              xlab = var, ylab = "Boarding")

      # Perform ANOVA
      anova_result <- summary(aov(boarding ~ data[[var]], data = data))
      cat("ANOVA p-value:", anova_result[[1]]$`Pr(>F)`[1], "\n")
    }}}
pairwise_analysis(capmetro_data)
```

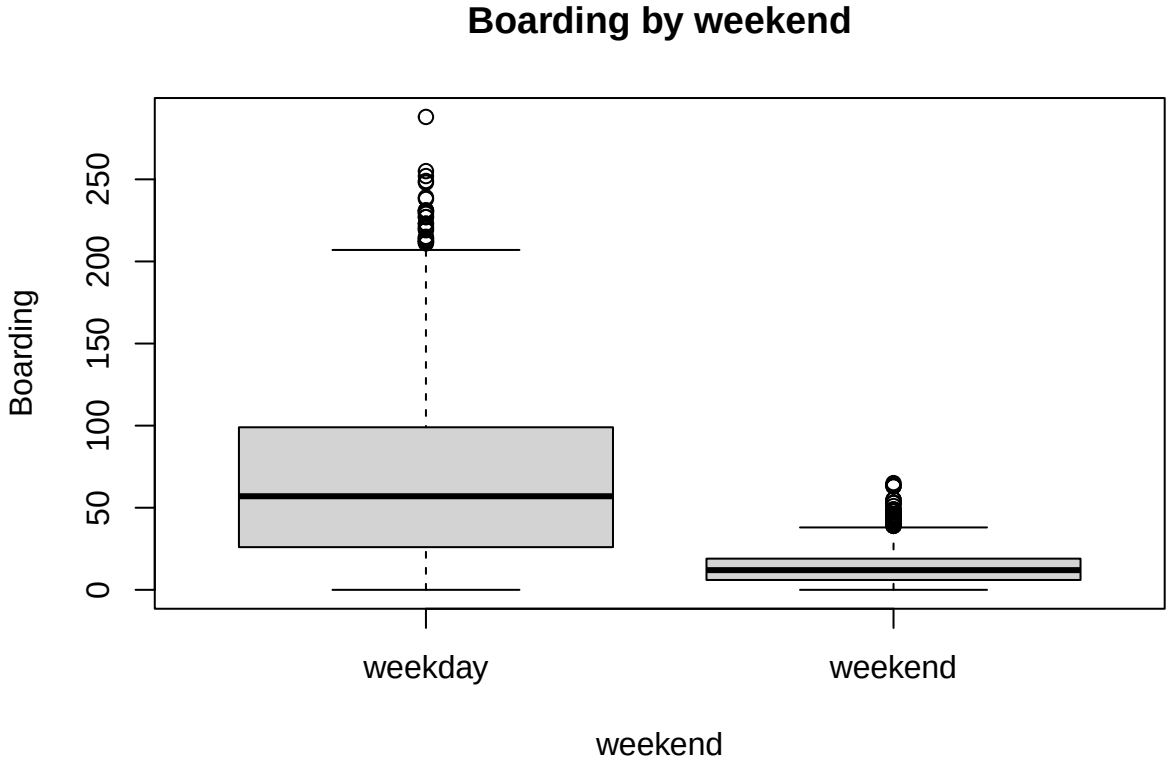
```
##
## Analyzing: hour_of_day
## Correlation with boarding: 0.352
```




```
##
## Analyzing: day_of_week
```




```
## ANOVA p-value: 0
##
## Analyzing: weekend
```




```
## ANOVA p-value: 0
```

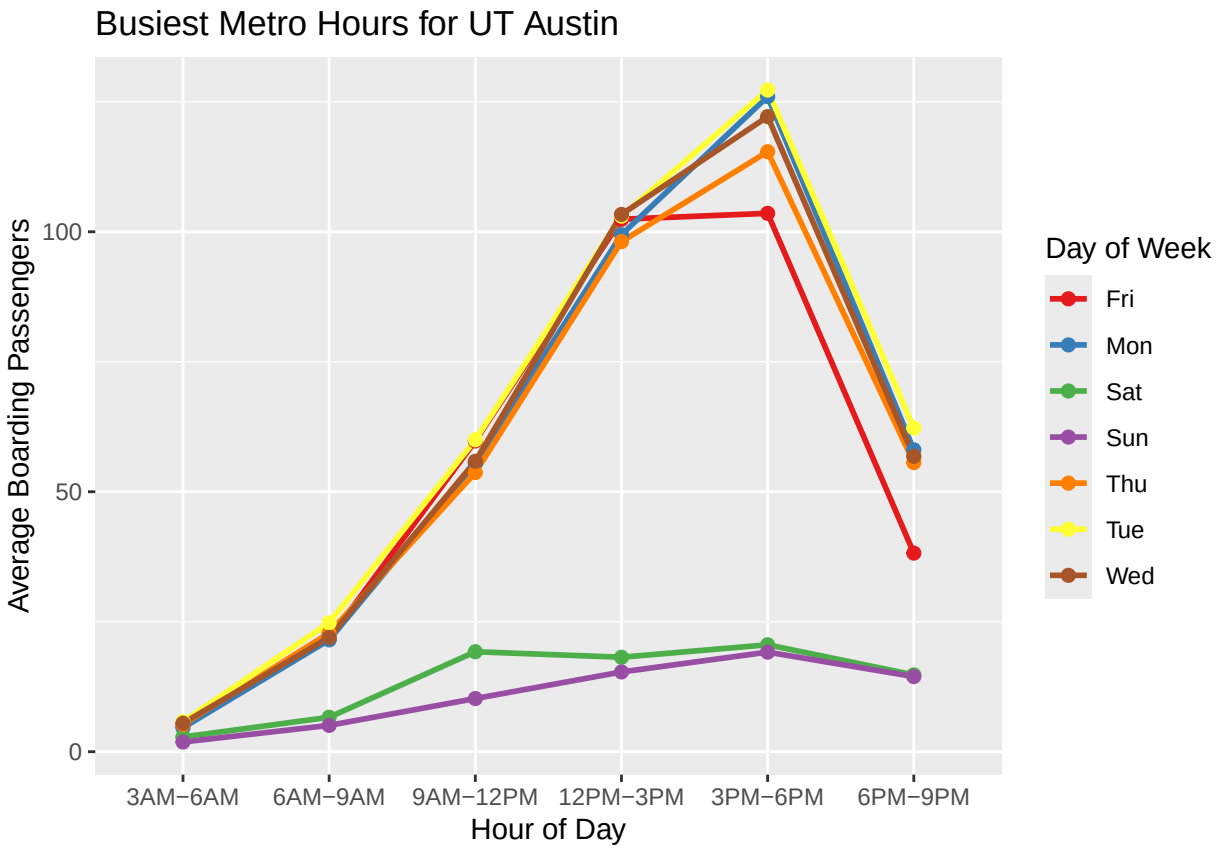
```r
cat("We can see a pretty clear relationship between what day and hour it is and
    how many people board. Now let's get rid of the weekend data and combine the
    days and hours data to see what the busiest boarding times all week are.")
```

```
## We can see a pretty clear relationship between what day and hour it is and
##     how many people board. Now let's get rid of the weekend data and combine the
##     days and hours data to see what the busiest boarding times all week are.
```

```r
# Create hour bins
capmetro_data <- capmetro_data %>%
  mutate(hour_bin = cut(hour_of_day,
                        breaks = seq(0, 24, by = 3),
                        labels = c("12AM-3AM", "3AM-6AM", "6AM-9AM", "9AM-12PM",
                                   "12PM-3PM", "3PM-6PM", "6PM-9PM", "9PM-12AM")))
```

```r
#plot data
capmetro_data %>%
  group_by(day_of_week, hour_bin) %>%
  summarise(avg_boarding = mean(boarding), .groups = "drop") %>%
  ggplot(aes(x = hour_bin, y = avg_boarding, color = day_of_week, group = day_of_week)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Busiest Metro Hours for UT Austin",
       x = "Hour of Day",
       y = "Average Boarding Passengers",
       color = "Day of Week")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
cat("From this chart, we can see that the busiest hours are from 3-6pm on Monday
    and Tuesday, with the number of people gradually dropping off the later you
    get into the week. Very few people ride on the weekends, to the point that
    more people ride the bus from 6-9am on Friday than 3-6pm on Sunday.")
```

```
## From this chart, we can see that the busiest hours are from 3-6pm on Monday
##     and Tuesday, with the number of people gradually dropping off the later you
##     get into the week. Very few people ride on the weekends, to the point that
##     more people ride the bus from 6-9am on Friday than 3-6pm on Sunday.
```